Machine Learning - Final Project

Spring 2024, Stephan Ohl

Overview

In this final project you should demonstrate your knowledge in ML and apply it to a basic machine learning problem. The project is subdivided into five parts: (1) loading and cleaning a data set, (2) dimensionality reduction, (3) training one or multiple models, (4) evaluate model or model set, (5) describing lessons learned.

Upload a zip-file on Canvas with this name: `ML_<name_group_head>.zip`. The file should contain (i) a file `GROUP.TXT` contain a list of all group members, (ii) one Jupyter notebook file containing all your code (including your explanations and descriptions for each project part), and (iii) the data set (<10 MB) that your code is operating on.

Part 1: Load and Clean Data

Use pandas or some other method to load your basic data set. You should use a data set that is not too big to keep all processing times reasonable for such an educational project. If your data set is too big, you can also subsample it.

You are free to choose your own data set. Keep in mind, however, that the data should be suitable for a classification or regression problem. We did not discuss, for instance, time series data in class. Some structures of data may requires more specialized models and algorithm. If you are unsure about where to get your data from, you should choose from an open data repository.

Open Data Repositories:

- UC Irvine ML Repository `archive.ics.uci.edu/ml/`
- Kaggle Datasets: `www.kaggle.com/datasets`

Make sure that your data loading mechanism creates a clean data set. You may want to remove instances with missing values or filter instances with wrong values.

Part 2: Dimensionality Reduction

Describe your data with a short text paragraph: describe the domain, number and type of attributes, number of instances, and anything else that seems to be relevant.

Then, use a dimensionality reduction technique to reduce the number of attributes in your data. For instance, you may want to apply PCA, Kernel PCA, or some Manifold Learning technique.

- https://scikit-learn.org/stable/modules/unsupervised_reduction.html
- https://scikit-learn.org/stable/modules/decomposition.html
- https://scikit-learn.org/stable/modules/manifold.html

Part 3: Training Model

Choose a ML model class and train a model instance. You may choose, for instance, the ML techniques that were covered in class: Decision Trees, Linear Regression, Ensemble Methods, or Neural Networks.

- https://scikit-learn.org/stable/supervised_learning.html
- https://www.tensorflow.org/overviewl

You should train your model on your dimensionality reduced training set. You may choose to train multiple different models and choose the best among them in the evaluation part. Make sure that your training set is not too big such that training is done in a reasonable time.

Part 4: Evaluate Model

Evaluate the prediction performance of your model. For a binary or multi-class classification model, print the confusion matrix for your model(s). For a regression problem, print, for instance, the RMSE. Also, measure model training and estimation times.

Part 5: Lessons Learned

Reflect on what you learned from implementing this project and what you might do differently if you had the time to start all over again. What was the most complicated part of the project and what took most of your time?